Proceedings of the International Conference "Underwater Acoustic Measurements: Technologies &Results"
Heraklion, Crete, Greece, 28th June – 1st July 2005

A SCIENTIFIC DATA MINING APPROACH TO MIDWATER MULTIBEAM ECHOSOUNDING FOR FISHERIES APPLICATIONS

Bart Buelens[a], Ray Williams[a], Arthur Sale[a], Tim Pauly[b]

[a]School of Computing, University of Tasmania, Hobart TAS 7001, Australia
[b]SonarData Pty Ltd, Hobart TAS 7001, Australia

Contact author: Bart Buelens, School of Computing, University of Tasmania,
Private Bag 100, Hobart TAS 7001, Australia
Fax: +61 3 6234 1822 - E-mail: bbuelens@utas.edu.au

Abstract: *Midwater acoustic backscatter measurements collected by multibeam sonar offer new opportunities and challenges for fisheries applications. A scientific data mining technique to handle midwater multibeam backscatter data is presented. Most of the earlier research on multibeam echosounding for fisheries has focused on the core basic technologies of multibeam transducers, the associated signal processing, and calibration. Some work has been done with post-processed data, but no systematic methodology for post-processing of midwater multibeam backscatter data has emerged. In this paper, the problem is placed in a data mining framework. A model inversion technique is utilized, by applying the inverse of an approximation to the multibeam echosounding model. The proposed approach leads to a data product consisting of a collection of midwater soundings. A multibeam midwater sounding is the equivalent of the standard multibeam soundings as obtained from hydrographic multibeam instruments. These soundings can be visualized directly, allowing for exploratory data analysis in a 3d or 4d interactive environment. A sounding is a measurement in space and time, and has associated attributes or features, such as the backscatter value. Other features can be tagged to the soundings, forming generalised data nodes. Advanced spatio-temporal data mining techniques can now be applied to this set of nodes. Some further clustering techniques are presented, clustering the soundings into groups representing coherent objects in the water column, or, more specifically, fish schools. Global properties of clusters can be derived from the individual feature tags of the soundings, thus allowing for classification of schools into classes of similar types. The latest developments of this research are presented.*

Keywords: *Fisheries acoustics, multibeam sonar, data analysis*

1. INTRODUCTION

Recent multibeam echosounder systems have the capability of recording backscatter data for the whole water column, not only for the seabed, as was previously the case with hydrographic multibeam systems. The collection of midwater backscatter data from such instruments is of particular interest to the fisheries acoustics research community. It is expected that much more detailed information can be derived from multibeam echosounder data compared with the conventional single beam echosounder systems. However, deriving useful information from multibeam data is not a trivial task. In this paper, the problem is formulated in a data mining context. Data mining is the process of deriving knowledge and information from large data sets which do not exhibit that information in a trivial manner [1].

Full water column multibeam backscatter data sets are much larger than single beam echosounder data sets, typically of the order of 100 to 200 times. The structure of the data differs from single beam data in that multibeam data has an additional spatial dimension, offering 3-dimensional measurements as opposed to the conventional 2-dimensional single beam data sets. Detecting objects such as fish schools is no longer a straightforward procedure. Exploring multibeam data is less intuitive, since the visualization of such data is a challenging computational task.

Automated techniques must be developed to analyse the data, enabling easy extraction of information and patterns from it, and providing more accessible data products. A data product is a representation of the data that highlights or clarifies particular aspects that were not obvious from the raw data. A data mining approach leading to useful data products for the fisheries acoustics research community is presented.

2. FULL WATER COLUMN MULTIBEAM ECHOSOUNDING

In fisheries acoustics applications, the typical echosounder being used for surveying is essentially a single beam system [2]. Variants include split beam and dual beam systems. In the last few years, multibeam systems that are capable of collecting data for the full water column have become available [3, 4]. Until then, multibeam sonars would gate out the midwater backscatter, and output data concerning the seabed only. Such systems are commonly used in hydrographic applications.

Providing access to midwater backscatter data is an interesting concept that is gaining in popularity, since it means a saving in costs of hardware and potentially software, and in survey costs: the same instrument can now be used simultaneously for hydrographic and for fisheries work. The multiple uses of multibeam data sets require a new approach to data analysis, which caters for all interested parties.

Hydrographers want to extract the same data from the data sets that they have traditionally collected. This includes bathymetry, and often also seabed backscatter data that can be used in seabed classification applications. Hydrographic multibeam systems extract such information, mostly using custom algorithms, and output it to disk for further use. The advantage of having access to the full water column data is that this extraction can now happen as a post-processing step, and be repeated with different parameters and algorithms if desired. This will enable better quality control through the optimal tuning of algorithm parameters. In addition, extra information can be extracted, which is not generally possible with standard hydrographic systems, for example information relating to the bottom echo return pulse. Such information will be of value to research concerning the seabed habitat.

Fisheries acousticians can utilize full water column multibeam data to their advantage since it provides an extra dimension and much larger sampling volume compared to single beam echosounder data. Fish schools can be measured in their full extent in three dimensions, reducing much of the uncertainty inherent to single beam echosounding. Information relating to fish school dimensions, shape and energy content is contained within the multibeam data sets [4, 5].

3. UNIFIED DATA ANALYSIS

Extracting useful information from multibeam data sets for all parties involved requires a novel approach to data analysis. Ideally, a single unified approach will lead to data products for fisheries, habitat mapping and hydrography.

A data product is a representation of some aspects of an underlying raw unprocessed data set, presented in such a form that it provides useful information to scientists. The process that leads from the raw data set to data products is referred to as data mining. When the data being mined is of a scientific nature, the term scientific data mining is commonly employed.

A data mining framework can be regarded as a derived data set with associated tools that facilitate further analysis and creation of final data products.

The next section describes the creation of an intermediate derived data representation which will enable the application of advanced analysis techniques.

4. MODEL INVERSION

In order to gain a better insight into the formative physical processes for multibeam data, and the effects of various parameters, a model was developed by the authors [6]. Formally,

$$\mathit{\Delta} = M(\Psi) \tag{1}$$

with

Ψ the underwater environment (input to the model),

M the model,

$\mathit{\Delta}$ the data (output of the model).

Ψ takes the form of a set of points, each point representing a point scatterer. Ψ is the input to a model *M* which includes a ray tracing model, as well as a model of the digital signal processor of a multibeam system, taking care of sampling and beamforming. The resulting data set $\mathit{\Delta}$ includes a sequence of acoustic images, as well as some associated meta-data, such as time tag and geographic location.

Eq. (1) can be interpreted as the measurements $\mathit{\Delta}$ being a representation of the underwater environment Ψ, through an acoustic multibeam system *M*. We assume that the model *M* is a faithful representation of a real multibeam system, and will treat it as such in what follows.

Data analysis algorithms have access to the measurements $\mathit{\Delta}$ only, and aim to describe properties of Ψ. It would be convenient to invert eq. (1), so obtaining direct access to Ψ:

$$\Psi = M^{1}(\mathit{\Delta})$$

Unfortunately, M is not invertible analytically. The model inversion technique adopted here is to approximate M with a function F for which the inverse F^{-1} exists and is available analytically.
Calculating

$F^{-1}(\Delta) = \hat{\Psi}$

leads to an approximation $\hat{\Psi}$ of Ψ. It is common practice in synthesis imaging applications such as medical ultrasound and astronomy, to approximate the measuring process by a convolution [7, 8]. Since multibeam echosounding is a synthesis imaging system, the model M –which is assumed to be a faithful representation of the multibeam echosounding process- is approximated by a convolution C. In other words, the choice for F is C, and

$$C^{-1}(\Delta) = \hat{\Psi} \tag{2}$$

C^{-1} is a deconvolution, which, when applied to the multibeam measurements, results in an approximation $\hat{\Psi}$ of the underwater environment Ψ being measured. More detail can be found in [6, 9].

5. A BASIS FOR DATA MINING

5.1. Midwater soundings

Eq. (2) provides the basic step towards establishing a context in which further advanced data mining techniques can be applied. Δ is the set of measurements, essentially a sequence of acoustic images. Since C^{-1} is a deconvolution, $\hat{\Psi}$ is a sequence of images as well, the deconvoluted versions of Δ. On the other hand, Ψ is a set of points. In fact, $\hat{\Psi}$ will be a sequence of images of points, and straightforward thresholding yields a set of points. It is convenient to imply this thresholding step in C^{-1}, so that $\hat{\Psi}$ can be considered as a set of points (rather than images of a set of points):

$\hat{\Psi} = \{\boldsymbol{s}_i\}$, $i = 1 \ldots N$, with N the cardinality of $\hat{\Psi}$.

The points $\boldsymbol{s}_i$ in $\hat{\Psi}$ can be thought of as *soundings*, maintaining consistency and analogy with hydrographic multibeam applications. It is important to note that, as in hydrography, a sounding is not necessarily a point scatterer in the water; rather, it is a conceptual measurement indicating the presence of a general object in the water, which could also be an extended or solid object, such as a fish school, or the seabed.

In their simplest basic form, soundings are spatio-temporal measurements of backscatter intensity. A sounding $\boldsymbol{s}$ can be written in terms of its components as

$$\boldsymbol{s} = (\boldsymbol{x}, t, b) \tag{3}$$

with $\boldsymbol{x}$ the spatial coordinates, t the time stamp and b the backscatter value.

5.2. Spatial clustering

As an example of a derived result, a simple spatial clustering technique is applied to $\hat{\Psi}$, clustering the set of soundings into disjoint subsets $\hat{\Psi}=\bigcup_j \hat{\Psi}_j$. The algorithm used is commonly referred to as DBSCAN [10]; it is a standard density based spatial clustering technique. The clusters of midwater soundings are subsequently internally joined into a mesh by a Delaunay triangulation [11] to form fish school objects and a seabed surface. See Fig. 1. The algorithm was applied to the spatial components $\boldsymbol{x}$ of the soundings $\boldsymbol{s}$ only.

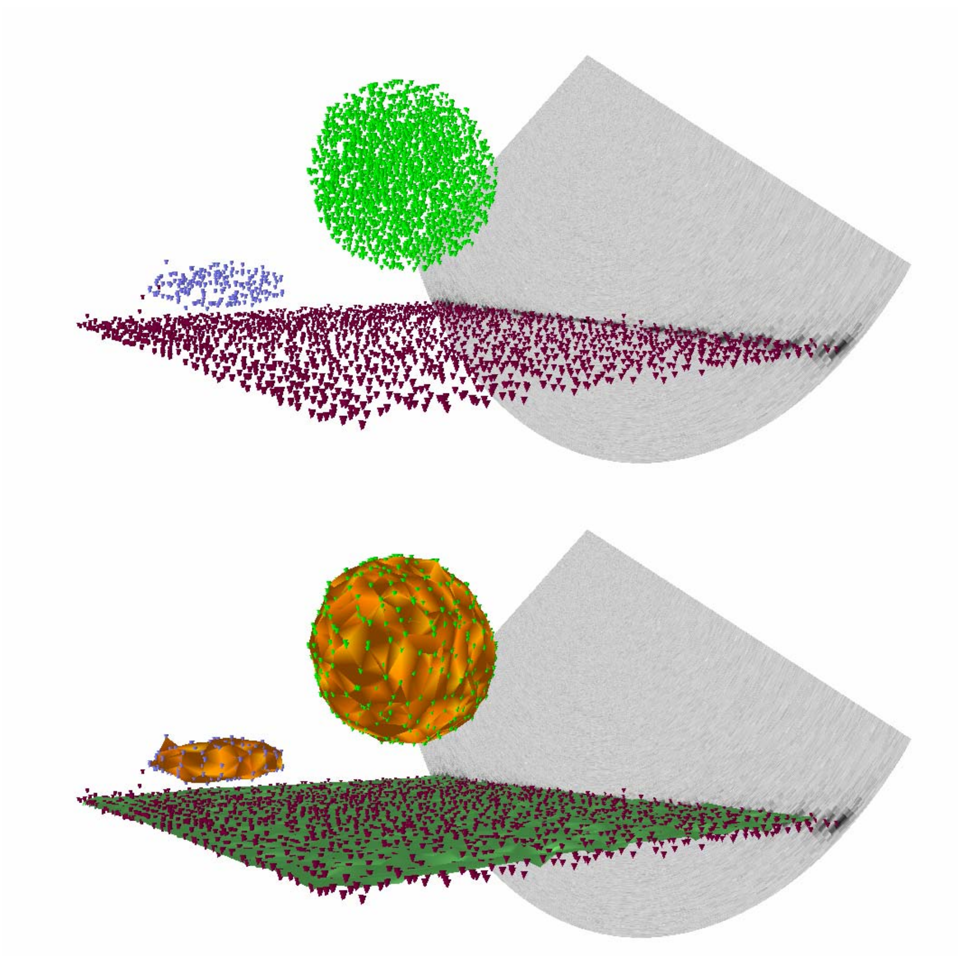

Fig. 1: (top) The soundings, colour-coded per cluster, (bottom) the result of the cluster-based object detection.

6. CONCLUSION AND FUTURE WORK

The example in the previous section demonstrates that the set $\hat{\Psi}$ offers a practical and useful basis for interesting further data analysis techniques.

The soundings as defined in (3) will be generalised, and extended to be *data nodes*. Apart from the spatio-temporal properties of soundings, data nodes will have a number of features associated with them, capturing a much wider and broader information content than just the soundings. A spatio-temporal data node $\boldsymbol{n}$ written in terms of its components: $\boldsymbol{n} = (\boldsymbol{x}, t, \boldsymbol{v})$. The vector $\boldsymbol{v}$ is a vector containing features of the node $\boldsymbol{n}$. The special case where there is only one feature and that feature is the backscatter value reduces $\boldsymbol{v}$ to b, as in (3). Current research is focussed on establishing $\boldsymbol{v}$, and on data mining techniques which utilize $\boldsymbol{v}$ for enhanced clustering, pattern recognition and classification.

7. ACKNOWLEDGEMENTS

SonarData Pty Ltd, Hobart, Tasmania, Australia are funding this research. Their continuing support is acknowledged. It is anticipated that some or all of the techniques and results presented in this paper will be incorporated in the SonarData software product range, in particular in the Echoview software package [5].

REFERENCES

[1] **N. Ramakrishnan and A. Y. Grama**, Mining scientific data, *Advances in Computers*, vol. 55, pp. 119-169, 2001.
[2] **D. N. McLennan and E. J. Simmonds**, *Fisheries Acoustics*, Chapman & Hall, 1991.
[3] **E. Hammerstad**, Advanced multibeam echosounder technology, *Sea Technology*, vol. 36, pp. 67-69, 1995.
[4] **F. Gerlotto, M. Soria, P. Fréon**, From two dimensions to three: the use of multibeam sonar for a new approach in fisheries acoustics, *Canadian Journal of Fisheries and Aquatic Science*, vol. 56, pp. 6-12, 1999.
[5] **SonarData Pty Ltd**, Echoview Documentation, Available on-line at *http://www.sonardata.com/webhelp/Echoview.htm*, 2005.
[6] **B. Buelens, R. Williams, T. Pauly, A. Sale**, Midwater acoustic modelling for multibeam sonar simulation, In *146th Meeting of the Acoustical Society of America*, Austin, Texas, 2003.
[7] **F. Lingvall**, A method of improving overall resolution in ultrasonic array imaging using spatio-temporal deconvolution, *Ultrasonics*, vol. 42, pp. 961-968, 2004.
[8] **J. L. Starck, E. Pantin, and F. Murtagh**, Deconvolution in astronomy: A review, *Publications of the Astronomical Society of the Pacific*, vol. 114, pp. 1051-1069, 2002.
[9] **B. Buelens, R. Williams, A. Sale, T. Pauly**, A Framework for Scientific Data Mining in Hydro-Acoustic Data Sets, In *Proceedings Second International Conference on Artificial Intelligence in Science and Technology (AISAT)*, Hobart, Australia, pp. 104-108, 2004.
[10] **M. Ester, H. Kriegel, J. Sander, X. Xu**, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, In *Proceedings of 2nd Interational Conference on Knowledge Discovery and Data Mining*, Portland, pp.226-231, 1996.
[11] **C. B. Barber, D.P. Dobkin, H.T. Huhdanpaa**, The Quickhull Algorithm for Convex Hulls, *ACM Transactions on Mathematical Software*, vol. 22, pp. 469-483, 1996.